Calibration of probabilistic age recognition

David A. van Leeuwen[1,3] *and Hasan Mohamad Basari*[2]

[1]CLST/CLS, Radboud University Nijmegen, The Netherlands
[2]Centre for Processing Speech and Images, KU Leuven, Belgium
[3]Netherlands Forensic Institute, Den Haag, The Netherlands

Abstract

The task in automatic age recognition in speech technology typically is one of regression, i.e., predicting the age of a speaker from his/her speech. In this paper we are interested in the probabilistic interpretation of the posterior distribution of the predicted age. We review a number of measures for assessing the probabilistic properties of the posterior distribution, and link these to detection theory, which is very well understood from the automatic speaker recognition literature. We show that the Gaussian posterior distributions predicted by least square support vector regression behave well, and that there is only a small room for improvement of their posterior distributions under the Gaussian assumption.

1. Introduction

Automatic age recognition in speech technology is one of the areas in speaker characterization [11]. The general task can be specified as predicting the age of a speaker from a sample of speech from that speaker. This can be carried out in a classification scenario [13, 2] using age groups, or by using regression. In this paper we focus on regression, i.e., predicting the age in years. More specifically, we are interested in the predicted uncertainty in doing this.

Rather that asking what the *age* of a speaker is, we want to know the *posterior probability distribution* over ages is. This information could for instance readily be used in investigative forensic scenarios where the demographic information of a population of potential suspects is known, but speaker models are not available. If such age information is to be combined with other evidence, it is important that the computed probabilities are *well calibrated*, so that the age information's importance is not over- or underestimated w.r.t. the other evidence.

This paper investigates performance measures for calibrated posterior distributions, and evaluates these for a LS-SVR age recognition system. The focus is on the task and measures in Section 2, then the experimental data and systems are briefly described in Section 3, and finally the calibration is assessed in Section 4, where we investigate if the LS-SVR posteriors can be re-calibrated with simple linear transformation.

2. Regression task

The task we are concerned with this paper is the prediction of the posterior distribution over ages a, given an utterance of speech x_i from a speaker S_i

$$p(a \mid x_i). \tag{1}$$

In a typical scenario, a MAP estimate of the age of the speaker, $\hat{a}$, can be computed from this as $\hat{a} = \arg\max p(a|x_i)$. The value $\hat{a}$ is usually the sole value reported in regression tasks such as age recognition, but here we are interested in the full probability density function (PDF). However, we will first discuss some evaluation measures for the point estimate $\hat{a}$.

2.1. Evaluation measures for $\hat{a}$

Perhaps the most intuitive error measure for prediction is the *mean absolute error*, measured on a set of N trials with true age a_i

$$E_{\mathrm{ma}} = \frac{1}{N}\sum_i |\hat{a}_i - a_i|. \tag{2}$$

An alternative to this is the square root of the mean of the squared differences,

$$E_{\mathrm{rms}} = \sqrt{\frac{1}{N}\sum_i (\hat{a}_i - a_i)^2}. \tag{3}$$

This measure puts more weight to larger deviations from the true age, and is often closer connected to the training objective in the machine learning technique employed. One could argue that age prediction from speech is a hard task, even for people, and that therefore the larger age differences are more interesting and should be weighted appropriately.

2.2. Evaluation measures for prediction uncertainty

In the PASCAL *Evaluating Predictive Uncertainty Challenge* (EPUC) and the follow up of that event a number of evaluation measures that indicate the goodness of the predicted PDF were proposed and discussed. We will briefly review two of these here now, and add the context of evaluation in automatic speaker recognition, where the measure of uncertainty has been developed quite extensively.

The official scoring function in EPUC is the loss function E_{nlpd}, *negative log predictive density*

$$E_{\mathrm{nlpd}} = -\frac{1}{N}\sum_i \log p(a_i \mid x_i). \tag{4}$$

This is a form of logarithmic scoring, which is also used in the measure C_{llr} in the evaluation of log-likelihood-ratio scores in speaker recognition [4] and Information Gain in weather forecasts [12]. The idea is that the system can gain performance by placing more probability mass close to the predicted age, but must reserve probability mass for other possible values of the age, or it runs the risk of an unbounded penalty. When interpreted as a negative score $S_{\mathrm{nlpd}} = -E_{\mathrm{nlpd}}$ this score is *strictly proper* [1].

This error measure was criticized by Kohonen and Suomela [8], for several reasons. The measure is only sensitive to the *local* PDF at a_i, and does not reward the predictor for having probability mass close to the actual value. It can even lead to misleadingly low error if assumptions about the resolution of the true age is known: if it is specified in integer years, for instance, the PDF could consist of infinitely narrow peaks around all integer values in the range 0–120, and the loss could be low without bounds. This particular case (of integer years) could be solved by requiring a discrete probability distribution, but Kohonen and Suomela instead propose another measure, the *continuous ranked probability score*, which is defined as the average squared difference of the cumulative distributions for the predicted $p(a \mid x_i)$ and the true age $\delta(a - a_i)$, respectively $P(A < a \mid x_i)$ and $u(a - a_i)$, the unit step function. When expressed as an error (lower meaning better prediction), the computation for a single trial is

$$E_{\text{crps}}(x_i) = \int_a \big(P(A < a \mid x_i) - u(a - a_i)\big)^2 w(a)\, da. \quad (5)$$

Here $w(a)$ is an arbitrary weight function, which we will take unity in this paper. For a set of trials this averages to

$$E_{\text{cprs}} = \frac{1}{N} \sum_i E_{\text{crps}}(x_i). \quad (6)$$

This E_{crps} is *distance sensitive* and *non-local* in the sense that probability mass closer to the true age lead to lower error. This scoring rule has been shown to be *proper* [9].

2.3. Evaluation of the cumulative distribution

Inspired by E_{crps} we may wonder how well the age estimator can be used as a detector for minimum age. By integration of the PDF we can define the posterior odds for speaker S_i having an age higher than a certain threshold t

$$\mathcal{O}^t_{\text{post}}(x_i) = \frac{P(A > t \mid x_i)}{P(A < t \mid x_i)}. \quad (7)$$

From this we can define a log-likelihood-ratio ℓ, by subtracting the log prior odds from the log posterior odds

$$\ell_t(x_i) = \log \frac{P(A > t \mid x_i)}{P(A < t \mid x_i)} - \log \frac{P(A > t)}{P(A < t)}. \quad (8)$$

The interpretation of this likelihood ratio is that of detection as in speaker recognition or forensic speaker comparison, and it can therefore be analyzed in terms of receiver operating characteristics (or Detection Error Trade-off). The calibration of these likelihoods can empirically be tested using C_{llr}, the cost of the log-likelihood-ratio. This error measure integrates both detection and calibration performance over a range of priors. C_{llr} is *strictly proper*, favoring properly calibrated likelihood ratios over other scores for all possible prior odds.

The C_{llr} for threshold t can be computed as

$$C^t_{\text{llr}} = -\frac{1}{2\log 2}\Big(\frac{1}{N_{a>t}} \sum_{i>} \log(1 + e^{-\ell_i}) + \frac{1}{N_{a<t}} \sum_{i<} \log(1 + e^{\ell_i})\Big), \quad (9)$$

where the sums are over trials with true age above and below the threshold t, respectively. C_{llr} penalizes under- and overconfidence in ℓ.

3. Age recognition system and data

We used two age recognition systems in this research, which both are based on least square support vector regression (LS-SVR) [14]. In the first system, labeled *GMM-SVR*, the feature vectors are UBM/GMM supervectors inspired by the well known GMM-SVM supervector approach in speaker recognition pioneered by Campbell [5] which stood at the basis of most competitive submissions to the NIST 2006 speaker recognition evaluation [3]. In this paper, the procedure of calculating GMM mean supervectors is very similar to the one described by Bocklet [2]. In the second system, labeled *i-vector-SVR*, i-vectors [7] extracted from the speech segments are used as features for the SVR. The procedure of extracting i-vectors in this paper is identical to the method described in [10].

We used two databases for evaluation of the age recognition systems, "N-Best" and "NIST SRE". The data in N-Best consisted of utterances from the N-Best Dutch speech recognition "Broadcast News" training data specification [16]. In total 555 wide-band utterances by 425 speaker were used from the Flemish parts of the database. This database was considered too small for a proper separation of training and testing, therefore we reverted to a 5-fold cross validation scheme. For each test fold of 111 segment, the remaining 444 segments were used for training and for determining the prior probabilities. Test speakers that occurred in the training folds were removed from the test, yielding 410 valid test utterances. The GMM-SVR age recognition model was trained in a gender-independent way.

The data from the NIST SRE database consisted of the telephone conversation speech segments from the NIST SRE 2010 core-core extended protocol for training the LS-SVR, amounting to 445 speakers in 5634 segments. For evaluation, the telephone conversation speech segments from SRE 2008 were used, with 1336 speakers in 3999 segments.

The LS-SVR [6] is capable of not only estimating an age $\hat{a}$, but giving a variance $\hat{\sigma}^2$ as well. Thus the PDF for a trial can be described as a normal distribution

$$p(a \mid x_i) = \mathcal{N}(a, \hat{a}_i, \hat{\sigma}_i) \equiv \frac{1}{\sqrt{2\pi}\hat{\sigma}_i} e^{-(a-\hat{a}_i)^2/2\hat{\sigma}_i^2}, \quad (10)$$

and the cumulative distribution as

$$P(A < a \mid x_i) = \Phi(a, \hat{a}_i, \hat{\sigma}_i) \equiv \frac{1}{2}\Big(1 + \text{erf}\,\frac{a - \hat{a}_i}{\sqrt{2}\hat{\sigma}_i}\Big). \quad (11)$$

The normal distribution has $\arg\max p(a \mid x_i) = E(p(a \mid x_i)) = \hat{a}$. The cumulative distribution is used for computing ℓ_i and E_{crps}—the latter can be carried out analytically in Φ but the expression is left out here for conciseness. For E_{nlpd}, the computation reduces to

$$E_{\text{nlpd}} = \frac{1}{N}\Big(\sum_i \log(\sqrt{2\pi}\hat{\sigma}_i) + (a_i - \hat{a}_i)^2/2\hat{\sigma}_i^2\Big), \quad (12)$$

which is minimum (for given $\hat{a}_i$) at $\hat{\sigma}_i^2 = (a_i - \hat{a}_i)^2$. Note, that when $\hat{\sigma} \to 0$, E_{crps} reduces to E_{ma}.

4. Experiments

The basic performance measures of the age recognition system are shown in Table 1. As a result of an evolving insight in both recognition technology and data bases, the GMM-SVR system was evaluated with the N-Best database in a cross validating scheme, while the i-vector-SVR system has been evaluated with

Table 1: Basic performance characteristics of the age recognition system. Lower numbers are better.

Prediction	Database	E_{ma}	E_{rms}	E_{nlpd}	E_{crps}
Prior	N-Best	8.79	10.90	3.80	6.21
GMM-SVR	N-Best	6.51	8.34	3.57	4.69
Prior	SRE male	9.28	11.50	3.90	6.55
i-vector-SVR	SRE male	7.61	9.88	3.89	5.79
Prior	SRE female	10.40	12.80	4.01	7.29
i-vector-SVR	SRE female	7.61	10.00	3.94	5.77

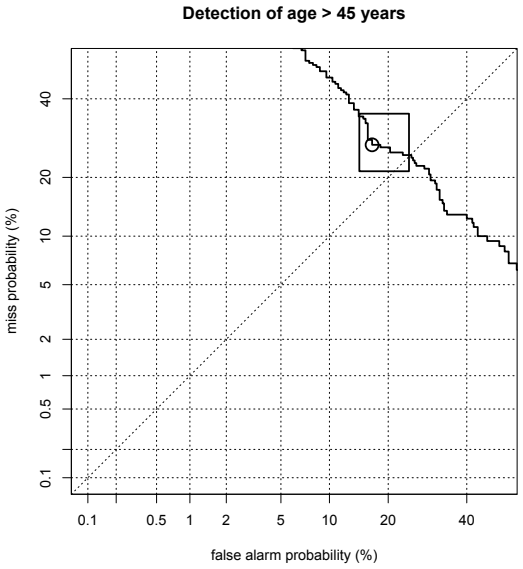

Figure 1: A Detection Error Trade-off for the age recognition, from integrated PDFs, for $t = 45$ years. The square indicates decisions taken based on $\hat{a}$, the circle 'minimum cost' for a prior 1/2 cost function.

the NIST SRE data set. The i-vector-SVR system is gender-dependent, hence we report separate results for the male and female portions of the NIST SRE data set. For each system, we report a method labeled "Prior" that simply predicts age $\hat{a}$ with variance $\hat{\sigma}^2$ based on the training data—for the GMM-SVR this is the 4/5 folds available for training.

One can observe that the measures for $\hat{a}$ show quite some improvement of the age recognition over just using a prior. For the evaluation of the predictive uncertainty E_{nlpd} seems quite pessimistic about the utility of the recognizer, where the non-local E_{crps} definitely shows evidence of a shift of probability mass towards the true age.

Next, we compute the log-likelihood-ratios for a detector with threshold age t, according to (8), where for the second term, the prior log odds, we simple use counts:

$$\log(\mathcal{O}^t_{\mathrm{prior}}) = \log \frac{\sum_i u(a_i - t)}{\sum_i u(t - a_i)} \tag{13}$$

where the summation is only over the training folds. In Fig. 1 we have plotted the resulting DET curve for $t = 45\,\mathrm{y}$. The fact that the ℓ computed from the posterior distributions gives rise to a reasonably shaped DET curve is not so surprising. This is basically true for any uni-modal posterior distribution in age. More interestingly, C_{llr} computed for these ℓ_i seems good over

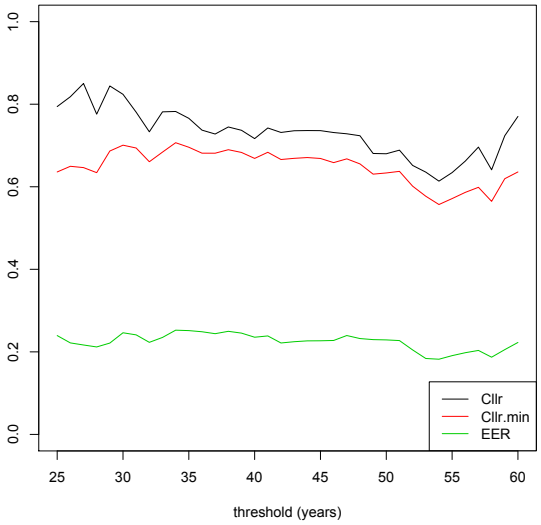

Figure 2: Detection and calibration metrics for a rage of threshold ages for the "GMM-SVR" system, C_{llr}, $C^{\mathrm{min}}_{\mathrm{llr}}$, and the Equal Error Rate $E_{=}$.

a wide range for age thresholds t. This can be appreciated from Fig. 2 where we have indicated a few key performance measures for the detection experiment, for a range of age thresholds. C_{llr} is below 1 for for all ages.

Apart from the Equal Error Rate $E_{=}$ and C_{llr}, a quantity $C^{\mathrm{min}}_{\mathrm{llr}}$ is plotted as well. This parameter has a clear meaning in a pure detection evaluation, indicating what C_{llr} could have been if the values ℓ_i had been optimally warped while maintaining their relative order [15]. Because our ℓ_i are a function $\ell(\hat{a}_i, \hat{\sigma}_i)$, the result of 'optimum warping' could be translated back to either one of these parameters or both. Therefore, the constriction 'while maintaining relative order' in ℓ becomes less meaningful to $\hat{a}$ and/or $\hat{\sigma}$. We find it more important, however, that C_{llr} does not deviate too much from $C^{\mathrm{min}}_{\mathrm{llr}}$, given the discrimination performance $E_{=}$.[1]

4.1. Calibration transformations

One of the questions we would like to address is if the posterior distributions (characterized by mean $\hat{a}_i$ and variance $\hat{\sigma}^2_i$ in our case) could have consistently been predicted better. With 'consistently' we mean in a global way, and not for each individual trial. Because in this research we concentrate on the distribution rather than a point estimate, we have investigated several transformations of the posterior that involve only σ_i. The general idea is that we choose some transformation $\hat{\sigma} \rightarrow f(\hat{\sigma})$ that minimizes a calibration-sensitive objective, and investigate the transformation parameters and performance measures. In Table 2 we give the results for a number of linear transformations. The gain in objective is relatively small, but the optimal parameters are reasonably consistent over all objective functions. We therefore conclude that for the GMM-SVR system tested on the N-Best data in a cross-validation setting the LS-SVR methods produces fairly well calibrated posterior

[1] Under assumptions of equal variance Gaussian distributions for "target" and "non-target" ℓ, $C^{\mathrm{min}}_{\mathrm{llr}}$ is quite well approximated by $C^{\mathrm{min}}_{\mathrm{llr}} \approx 1 - (2E_{=} - 1)^2$.

Table 2: Several calibration transformations for $\hat{\sigma}$ the "GMM-SVR" system evaluated on N-Best data. $\langle C^t_{\mathrm{llr}} \rangle$ is the average of C^t_{llr} over a range of ages from 25–60 years.

$f(\hat{\sigma})$	obj.	a, b	E_{crps}	E_{nlpd}	$\langle C^t_{\mathrm{llr}} \rangle$
identity	–	–	4.687	3.578	0.740
b	E_{crps}	$b = 8.17$	4.678	3.540	0.719
b	E_{nlpd}	$b = 8.34$	4.679	3.540	0.717
b	$\langle C^t_{\mathrm{llr}} \rangle$	$b = 8.79$	4.685	3.542	0.715
$a\hat{\sigma}$	E_{crps}	$a = 1.13$	4.671	3.544	0.715
$a\hat{\sigma}$	E_{nlpd}	$a = 1.19$	4.674	3.541	0.712
$a\hat{\sigma}$	$\langle C^t_{\mathrm{llr}} \rangle$	$a = 1.22$	4.676	3.541	0.712
$a\hat{\sigma} + b$	E_{crps}	$0.63, 3.65$	4.657	3.524	0.703
$a\hat{\sigma} + b$	E_{nlpd}	$0.55, 4.35$	4.658	3.524	0.703
$a\hat{\sigma} + b$	$\langle C^t_{\mathrm{llr}} \rangle$	$0.61, 4.16$	4.659	3.525	0.702

distributions—although a global σ for all trials can perform better, if it is chosen in retrospect.

For the i-vector-SVR system evaluated using the NIST SRE databases we observe similar behaviour, but error rates are higher (cf. Table 1), the DET shows worse detection performance (with $E_{=} \approx 26\,\%$ for female and $31\,\%$ for male, on average). For some age thresholds at the edge of the scale $C^t_{\mathrm{llr}} > 1$, and, in line with the above, we observe slightly better calibration if an optimal global σ is chosen than using the per-trial $\hat{\sigma}_i$ from the SVR.

5. Conclusions

We have investigated the calibration of the predictive distributions of an age recognition system. The three calibration measures E_{nlpd}, E_{crps} and C_{llr} all indicate that the calibration is reasonable good, as no exceptionally high values are observed. Further, they all seem to behave similarly when optimizing the width of the distribution. Results on the NIST SRE database showed generally less performance than on N-Best, both in accuracy in the point estimate $\hat{a}$ and calibration/detection performance. Since the i-vector features are expected to generally work better than GMM supervectors, this may be attributed to the telephone bandwidth, and the age distribution of SRE data which is more skewed towards younger people (a median of 34 years vs. 43 for N-Best).

In this research we have not reported on transformations of $\hat{a}_i$ which could also be considered aspects of calibration for a regression problem. We see this as an accuracy issue that obviously lies at the heart of the regression problem, but that does not reveal the probabilistic character of the predictive distribution.

6. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Frame work Programme (FP7/2007–2013) under grant agreement no. 238803. We would like to thank Niko Brümmer for a helpful discussion.

7. References

[1] Jose M. Bernardo. Expected information as expected utility. *Annals of Statistics*, 7(3):686–690, 1979.

[2] T. Bocklet, A. Maier, and E. Noth. Age determination of children in preschool and primary school age with GMM-based supervectors and support vector machines/regression. In *Proc. 11th Int. Conf. Text, Speech and Dialogue*, pages 253–260, 2008.

[3] Niko Brümmer, Lukáš Burget, Jan Černocký, Ondřey Glembek, František Grezl, Martin Karafiát, Pavel Matějka, David A. van Leeuwen, Petr Schwarz, and Albert Strassheim. Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. *IEEE Transactions on Speech, Audio and Language Processing*, 15(7):2072–2084, 2007.

[4] Niko Brümmer and Johan du Preez. Application-independent evaluation of speaker detection. *Computer Speech and Language*, 20:230–275, 2006.

[5] William Campbell, Douglas Sturim, and Douglas Reynolds. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5):308–311, 2006.

[6] K. De Brabanter, J. De Brabanter, J.A.K. Suykens, and B. De Moor. Approximate confidence and prediction intervals for least squares support vector regression. *IEEE Transactions on Neural Networks*, 22(1):110–120, January 2011.

[7] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):788–798, May 2011.

[8] Jukka Kohonen and Jukka Suomela. Lessons learned in the challenge: Making predictions and scoring them. volume 3944 of *Lecture Notes in Computer Science*, pages 95–116. Springer Berlin / Heidelberg, 2006.

[9] James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976.

[10] Mitchell McLaren and David A. van Leeuwen. Source-normalised LDA for robust speaker recognition using i-vectors from multiple speech sources. *IEEE Transactions on Audio, Speech and Language Processing*, 20(3):755—766, March 2012.

[11] Christian Müller, editor. *Speaker Classification I*, volume 4343 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2007.

[12] Riccardo Peirolo. Information gain as a score for probabilistic forecasts. *Meteorological Applications*, 18:9–17, 2011.

[13] Susanne Schötz. Acoustic analysis of adult speaker age. In Christian Müller, editor, *Speaker Classification I*, volume 4343 of *Lecture Notes in Computer Science*, pages 88–107. Springer Berlin / Heidelberg, 2007.

[14] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific Pub. Co., Singapore, 2002.

[15] David A. van Leeuwen and Niko Brümmer. An introduction to application-independent evaluation of speaker recognition systems. In Christian Müller, editor, *Speaker Classification*, volume 4343 of *Lecture Notes in Computer Science / Artificial Intelligence*. Springer, 2007.

[16] David A. van Leeuwen, Judith Kessens, Eric Sanders, and Henk van den Heuvel. Results of the N-Best 2008 Dutch speech recognition evaluation. In *Proc. Interspeech*, pages 2571–2574, Brighton, September 2009. ISCA.